Stylometry research using syntax-based features and Machine Learning techniques

Kim Luyckx (promotor Walter Daelemans)
CNTS Language Technology Group

September 25, 2006
CIL Research Contact Day

Universiteit Antwerpen

Outline

- PhD research
- Stylometry?
- Methodology
- Exploratory experiments
  - Corpus
  - Results & discussion
  - Conclusions
- PhD: Expected results

PhD Research (begin 2007 – end 2010)

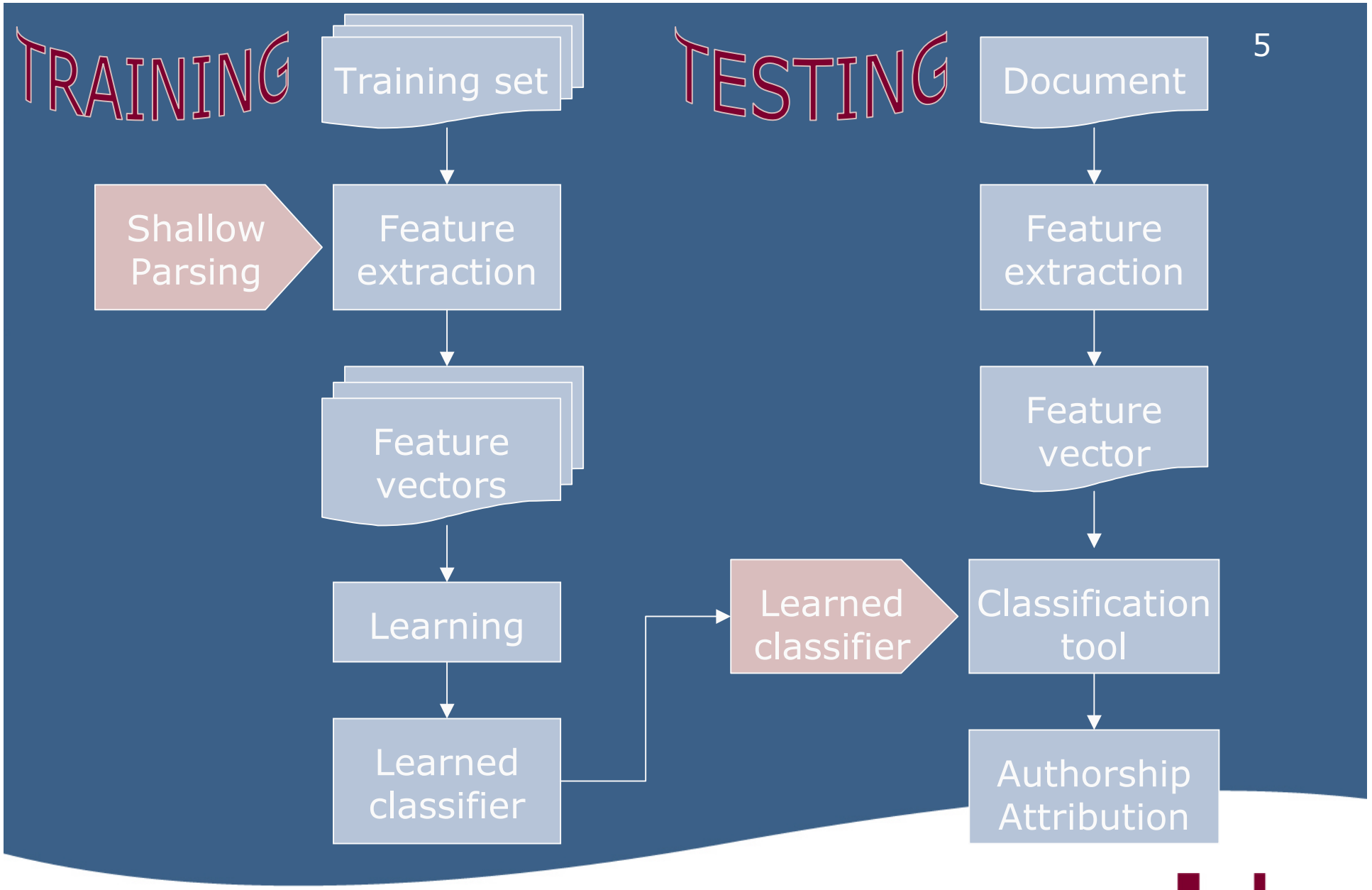
- Technical & methodological infrastructure for applied stylometry for Dutch

- Development of tools
  - Corpora
  - Benchmarks
  - Software for linguistic analysis
- Main facets
  - Automatic linguistic analysis
  - (un)superised learning
  - Evaluation

Stylometry?

- 'Stylistic genome' (author, period, genre, register)
  - Style characteristics (invariants): lexical, morphological & syntactic
  - Fiction vs. non-fiction
  - Sex and age groups
- Applications
  - Disputed authorship
  - Historical changes in style – Document dating
  - Gender detection
  - Forensic linguistics
  - Plagiarism detection (students, internet, programs)

Methodology


TRAINING
Training set
Shallow Parsing
Feature extraction
Feature vectors
Learning
Learned classifier
TESTING
Document
Feature extraction
Feature vector
Learned classifier
Classification tool
Authorship Attribution

Feature Extraction

- Memory-Based Shallow Parsing
  - Part-of-speech tagging

[$NP_1^{Subject}$ The/DT current/JJ advances/NNS $NP_1^{Subject}$]
{PNP [P in/IN P] [NP shallow/NN parsing//NN NP] PNP}
[VP_1 allow/VB VP_1] [$NP_2^{Subject}$ us/PRP $NP_2^{Subject}$]
[VP_2 to/TO use/VB VP_2] [NP_2^{Object} insights/NNS NP_2^{Object}]
{PNP [P from/IN P] [NP this/DT field/NN NP] PNP}
{PNP [P in/IN P] [NP stylometry//NN research/NN NP] PNP} ./.

Feature Extraction

- Memory-Based Shallow Parsing
  - Part-of-speech tagging
  - Chunking

[NP_1^Subject The/DT current/JJ advances/NNS NP_1^Subject]
{PNP [P in/IN P] [NP shallow/NN parsing//NN NP] PNP}
[VP_1 allow/VB VP_1] [NP_2^Subject us/PRP NP_2^Subject]
[VP_2 to/TO use/VB VP_2] [NP_2^Object insights/NNS NP_2^Object]
{PNP [P from/IN P] [NP this/DT field/NN NP] PNP}
{PNP [P in/IN P] [NP stylometry//NN research/NN NP] PNP} ./.

Feature Extraction

- Memory-Based Shallow Parsing
  - Part-of-speech tagging
  - Chunking
  - Identification of syntactic relations

[NP_1^Subject The/DT current/JJ advances/NNS NP_1^Subject]
{PNP [P in/IN P] [NP shallow/NN parsing//NN NP] PNP}
[VP_1 allow/VB VP_1] [NP_2^Subject us/PRP NP_2^Subject]
[VP_2 to/TO use/VB VP_2] [NP_2^Object insights/NNS NP_2^Object]
{PNP [P from/IN P] [NP this/DT field/NN NP] PNP}
{PNP [P in/IN P] [NP stylometry//NN research/NN NP] PNP} ./.

Feature Extraction

- Memory-Based Shallow Parsing
  - Part-of-speech tagging
  - Chunking
  - Identification of syntactic relations

- Feature vectors (f_1,f_2,...,f_n,class)
  - One vector per document
  - Comma-separated features
  - Class label

Learning/Classification

- Training and test phase
- Machine Learning algorithms
  - WEKA (Naive Bayes, Decision Trees, kNN, Neural Networks)
  - TiMBL: weighted kNN
- Ensemble methods (bagging & boosting)
- Feature weighting & selection

Exploratory experiments in Authorship Attribtion

Corpus

- Texts from *De Standaard*
- National politics section
- Similar genre and topics
- Average document length: ± 600 words

Class	Training corpus	# words	Test corpus	# words
A (Anja Otte)	100 articles	57,682	34 articles	20,739
B (Bart Brinckman)	100 articles	54,479	34 articles	25,684
O (The Others)	100 articles	62,531	32 articles	21,871

Possible markers of style

- Type-token ratio
- Word length
- Readability (Flesch-Kincaid metric)

$$206.835 - 1.015\left(\frac{\#\,words}{\#\,sentences}\right) - 84.6\left(\frac{\#\,syllables}{\#\,words}\right)$$

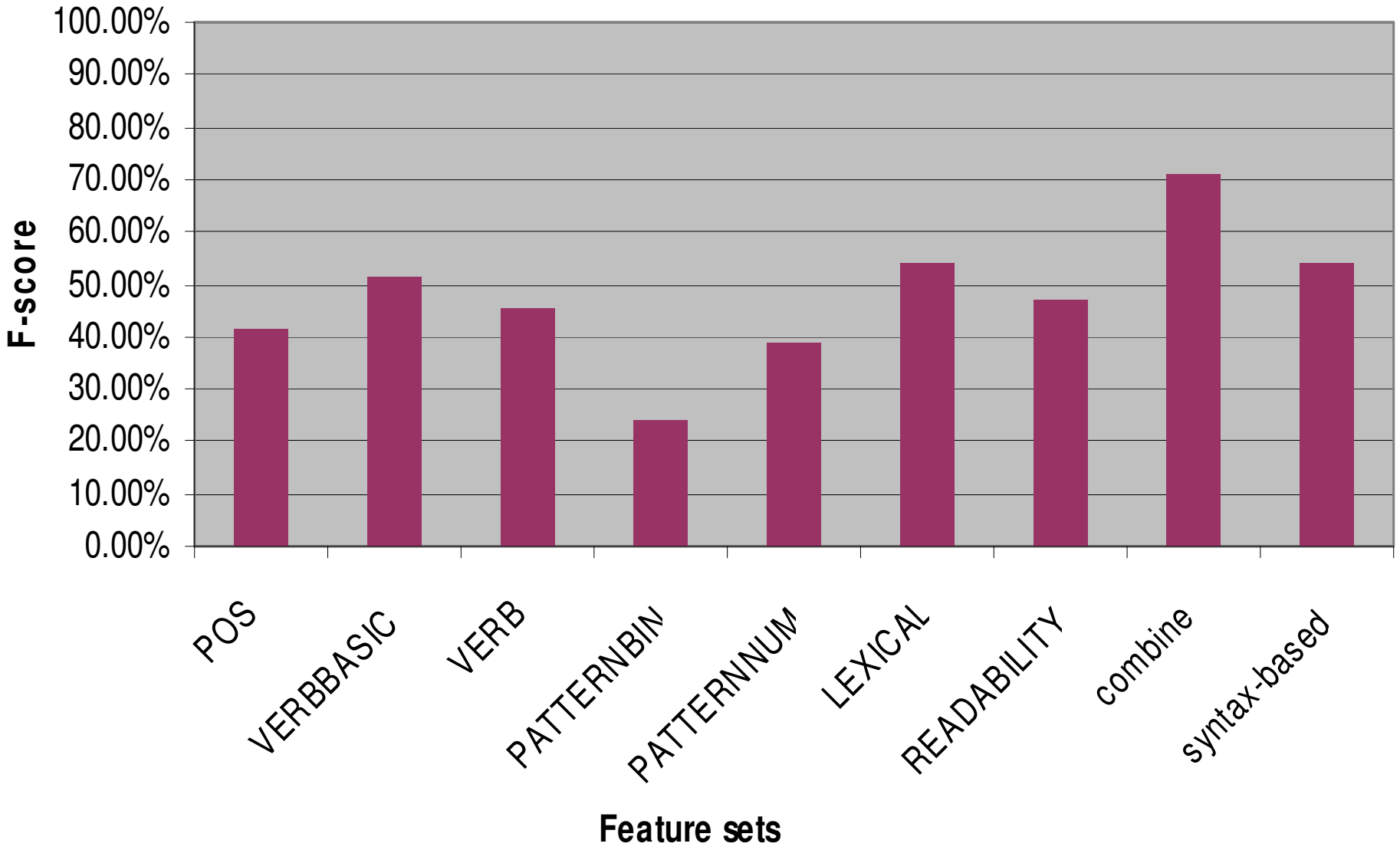
- Distribution of parts-of-speech & chunks
- Distribution of frequent function words
- NP and VP chunk internal variation

Results

- 3 authors: A vs. B vs. O
- 2 authors: A vs. B

- TiMBL & WEKA NNet
- F-score: weighted harmonic mean of precision & recall

$$F_\beta = \frac{(\beta^2 + 1) * \pi_i * \rho_i}{\beta^2 * \pi_i * \rho_i}$$


Performance on three author classes (TiMBL)
F-score
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
POS
VERBBASIC
VERB
PATTERNBIN
PATTERNNUM
LEXICAL
READABILITY
combine
syntax-based
Feature sets

Conclusions I: A vs. B vs. O

- Best feature sets: *verbbasic*: 52%
  lexical: 54%
- All feature sets *combine*d: 71% F-score
- *Syntax-based* features: 54% F-score

- WEKA NNet: best feature set: *lexical* (63.63%)
  combine 57.40%
  syntax-based 54.87%


Performance on two author classes (TiMBL)
F-score
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
POS
VERBBASIC
VERB
PATTERNBIN
PATTERNNUM
LEXICAL
READABILITY
combine
syntax-based
Feature sets

Conclusions II: A vs. B

- Best feature sets: *patternbin: 65%*
  lexical: 73%
- All features *combine*d: 81% F-score
- *Syntax-based* features: 66% F-score

- WEKA NNet: *combine* 65.25%
  syntax-based 64.45%

Conclusions

- Syntax-based, lexical and token-level feeatures are able to successfully tackle Authorship Attribution problems
- Syntax-based features perform equally well or sometimes even better!

PhD research: Expected results

- Operationalization of methodology (software package)
  - Text analysis tools
  - Tools for style feature extraction by means of Machine Learning
- Corpora for future research
- Answer fundamental research questions
  - Methodology vs. (non-)constant theme & register
  - Methodology vs. manual style analysis
  - Predictive power of different syntactic features
  - Applicability to authorship attribution & gender identification

Contact

- Project URL
  http://www.cnts.ua.ac.be/~kim/Stylometry.html
- Kim Luyckx
  - kim.luyckx@ua.ac.be
  - http://www.cnts.ua.ac.be/~kim
- Walter Daelemans
  - walter.daelemans@ua.ac.be
  - http://www.cnts.ua.ac.be/~walter

Questions?